TuBA: Tunable Biclustering Algorithm Reveals Clinically Relevant Tumor Transcriptional Profiles in Breast Cancer

Amartya Singh[1,2], Gyan Bhanot[1,2,3], Hossein Khiabanian[1,2,3,4]

[1] Department of Physics and Astronomy, Rutgers University, Piscataway, NJ, USA.

[2] Center for Systems and Computational Biology, Rutgers Cancer Institute, Rutgers University, New Brunswick, NJ, USA.

[3] Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, NJ, USA.

[4] Department of Pathology and Laboratory Medicine, Rutgers Robert Wood Johnson Medical School, Rutgers University, New Brunswick, NJ, USA.

Corresponding addresses

Hossein Khiabanian, PhD

Assistant Professor of Pathology in Medical Informatics

Rutgers Cancer Institute of New Jersey, Rutgers University

195 Little Albany Street, New Brunswick, NJ, 08903-2681

Phone: (732) 235 7554; E-mail: h.khiabanian@rutgers.edu

ABSTRACT

Clustering approaches that rely on a large number of variables, such as expression levels of thousands of genes, are often not well adapted to address the complexity and heterogeneity of tumors where small sets of genes may drive multiple cellular processes associated with carcinogenesis. Biclustering algorithms that perform local clustering on subsets of genes and conditions help address this problem. We propose a Tunable Biclustering Algorithm (TuBA) based on a novel pairwise proximity measure among gene pairs, which examines the relationship of samples at the extremes of genes' expression profiles to identify similarly altered signatures. The identified pairwise associations are illustrated graphically with nodes and edges representing the genes and the shared samples, respectively. Robust biclusters are then identified in these graphs iteratively. The consistency of TuBA's predictions was tested by comparing biclusters in 3,940 Breast Invasive Carcinoma (BRCA) samples from three independent sources, which employed different technologies for gene expression analysis (RNAseq and Microarray). Over 60% of the biclusters identified independently in each dataset had significant agreement among associated genes as well as similar clinical implications. About 50% of the biclusters were enriched in the ER-/HER2- (or basal-like) subtype, while more than 50% were associated with transcriptionally active copy number changes. Biclusters associated with gene expression patterns in non-malignant tissue were also found in tumor specimens. Overall, our method identifies a multitude of altered transcriptional profiles associated with the tremendous heterogeneity of diseased states in breast cancer, both within and across tumor subtypes, which is an important advance in understanding disease heterogeneity, and a necessary first step in individualized therapy.

INTRODUCTION

The first step in organizing and analyzing high-throughput gene expression datasets is to group together (cluster) similar biological variables (e.g. gene expression) or clinical conditions (e.g. patient samples) based on some mathematical measure of similarity. Since a priori knowledge about both the relevant variables and conditions is usually limited, clustering is often performed in an unsupervised manner, by grouping variables or conditions [1-5]. This global clustering is not necessarily an optimal approach for complex and heterogeneous cancers where, even within a clinical cancer type in a given tissue, only a few genes in subsets of patients are dysregulated in cellular processes associated with carcinogenesis. Global clustering is further confounded by the fact that a single gene can regulate and participate in multiple pathways in different conditions.

To address these concerns, a variety of local biclustering algorithms have been proposed that satisfy the following requirements: 1) a cluster of genes is defined with respect to only a subset of conditions and vice versa, and 2) the clusters are not exclusive and/or exhaustive – i.e. a gene/condition may belong to more than one cluster or to none at all [6-9]. Based on the type of biclusters and the mathematical formulation used to discover them, Oghabian *et al* [10] categorized biclustering techniques into four classes: (i) correlation maximization methods that identify subsets of genes and samples where the expression values of genes (or samples) is highly correlated across samples (or genes) [6]; (ii) variance minimization methods that identify biclusters where the expression values have low variance among the selected genes or conditions or both [11]; (iii) two-way clustering methods that iteratively perform one-way clustering on the genes and samples [12], and (iv) probabilistic and generative methods that employ stochastic approaches to discover genes (or samples) that are similarly expressed in subsets of samples (or genes) [13] [14].

Most of these biclustering methods rely on the actual gene expression values to perform their analysis. However, due to experimental noise or variability in thresholds in the expression levels of aberrantly expressed genes in diseased states, such an approach can miss alterations affecting only small groups of tumors. In this paper, we introduce a graph-based method called Tunable Biclustering Algorithm or TuBA, based on a novel measure of proximity that leverages the size of the cancer gene expression datasets to preferentially identify aberrantly co-expressed genes and pathways within subsets of patients. A key feature of the proximity measure used in TuBA is that it does not rely explicitly on the actual gene expression values. We demonstrate the utility of TuBA by applying it to three large, independent cohorts of breast invasive carcinoma (BRCA) encompassing 3,940 patients. In addition to detecting known pathways and subtypes associated with breast cancer, TuBA was able to uncover several novel sets of co-expressed genes across subtypes that may be relevant as biomarkers for therapeutic identification and intervention.

METHODS

Proximity measure. TuBA's proximity measure is based on the reasonable hypothesis that if a cellular mechanism is affected in a set of tumors, genes relevant to the mechanism should co-exhibit similar up or down regulation in a significant fraction of the tumors in the set. Consequently, TuBA examines all gene pairs to identify those that share a significant number of samples in their top/bottom percentile sets (**Fig. 1**). This measure is distinct from the usual differential expression analyses in two ways: (i) it is unsupervised and does not rely on a pre-selection of subtypes; (ii) it identifies co-aberrant gene patterns rather than individual differentially expressed genes. Since no penalty is imposed on relative changes in ranks of samples in the outlier percentile sets,

this measure is more robust against noise compared to other proximity measures, such as the Spearman's Rank correlation.

Graph-based Algorithm. For each gene, TuBA identifies samples in a pre-determined upper or lower percentile set. Pairwise comparison of these percentile sets using the Fisher's exact test identifies gene pairs that share a statistically significant number of samples. Each significant gene pair is illustrated graphically as pairs of nodes connected by an edge that represents the shared samples in their percentile sets. From the complete set of these pairwise graphs, robust co-aberrant signatures are established using the following iterative process (**Fig. 2**):

1. The graph is pruned such that its elementary units are complete subgraphs (cliques) of size 3 (triangles).
2. The largest clique (i.e. the seed) in the pruned graph is identified using the Bron-Kerbosch algorithm [15]. In cases where the largest clique is not unique, the union of all equally large cliques with a non-zero intersection of their nodes is designated as the seed.
3. The graph is trimmed by removing all the edges that contain any of the nodes in the seed in step 2. This step significantly reduces the computation time required to identify all the robust cliques in the graph.
4. Steps 2 and 3 are repeated till the graph has no elementary units left.
5. The seeds identified in steps 1-4 are exclusive in their gene sets, i.e., no two seeds share a common gene. To create the bicluster, the seeds are reintroduced sequentially into the original pruned graph from step 2, and nodes that share edges with at least two nodes in each seed are identified and added to the seed. The resulting graphs are the final biclusters obtained by TuBA.

TuBA's biclusters consist of sets of genes (nodes) that co-exhibit elevated or suppressed expression levels in subsets of samples (edges). Gene enrichment analysis of the gene sets can be used to identify their functional relevance and sample enrichment analysis can elucidate potential clinical subtypes and mechanisms. Furthermore, within each bicluster, genes can be assigned degrees, which are the total number of edges that connect them to other genes in the graph. Genes with higher degrees exhibit aberrant co-expression with other genes in a higher number of samples in the bicluster and hence are potential candidate driver genes.

Tuning TuBA. TuBA has two adjustable parameters:

1. The Percentile cutoff: The top or bottom percentile set gene expression threshold based on which overlap of sample sets is assessed. This parameter controls the number of samples considered for comparison.
2. The Overlap significance cutoff: The p-value threshold used to assess significance of the overlap of samples between percentile sets. This parameter controls the minimum number of samples that must be shared in percentile sets for their association to be considered significant.

For a given dataset, the choice of these two parameters determines the number as well as the composition of the final biclusters. The choice for these parameters depends on the heterogeneity in genomic alterations in tumors as well as the differences in their prevalence. Therefore, it is best to interpret the parameters as "knobs" that can be tuned to probe different levels of heterogeneity in the population (**Supplementary Methods**). TuBA is open-sourced and available as R scripts at https://github.com/KhiabanianLab.

Datasets. We applied TuBA to three independent BRCA datasets that emploied distinct methods for measuring transcript levels: (i) TCGA RNASeq gene expression dataset using the Illumina HiSeq 2000 RNA sequencing platform, (ii) METABRIC gene

expression dataset using the Illumina HT-12 v3 microarray platform, and (iii) six cohorts with gene expression data from GEO using the Affymetrix HGU133A microarray platform. To compare results among datasets, we applied TuBA to only their common gene sets. For clinical association analysis, we prepared two separate datasets for patients with known recurrence free survival (RFS) status (908 patients) and patients with known PAM50 subtype annotation (522 patients) respectively. Henceforth, we will refer to TCGA RFS, METABRIC RFS and GEO RFS datasets simply as TCGA, METABRIC and GEO, respectively, and PAM50 datasets are indicated specifically.

1) TCGA – BRCA: The $\log_2(x+1)$ transformed RSEM normalized counts of Level 3 data (2016-08-16 version), the clinical data (including relapse status and PAM50 subtype annotation from the 2012 Nature study [16]) (2016-04-27 version), and gene-level copy number variation (CNV) data, as estimated by GISTIC2 [17] (2016-08-16 version) were downloaded from the UCSC Xena Portal (http://xena.ucsc.edu). Genes with zero expression in all samples as well as the samples with NAs for any gene were removed from the analysis.

2) METABRIC: Normalized gene expression data, the clinical file, and the copy number file for the METABRIC study were downloaded from the cBioPortal (http://www.cbioportal.org) on 2017-05-14 [18, 19]. Gene expression dataset of 1,970 samples that had both relapse status and PAM50 subtype annotation were used in this study.

3) GEO: MAS5 normalized gene expression data and the clinical data with relapse status were downloaded from [20] on 2017-05-10. The dataset comprises samples from six independent cohorts. After processing, our gene expression dataset consisted of 1,062 patients with relapse status.

4) ESTIMATE: Scores for the level of stromal cells present and the infiltration level of immune cells in tumor tissues for 906 out of 908 samples for the TCGA – BRCA RNA-SeqV2 dataset using the ESTIMATE algorithm were downloaded from http://bioinformatics.mdanderson.org/estimate on 2017-10-12.

5) GTEx: RNASeq raw counts data from the Genotype-Tissue Expression (GTEx) portal (www.gtexportal.org/home/) was downloaded on 2017-06-15. The dataset comprised of all tissue samples currently available in the GTEx database. The 214 breast tissue samples were identified and normalized using the DESeq package in R [21].

Statistical Analysis. All computations were performed with R 3.3.0 [22]. The igraph package [23] was used to perform network/graph computations with some data summary functions performed using the plyr package [24]. The data.table package was used to handle data files, and the ggplot2 package was used to make plots [25]. GeneSCF [26] was used to perform gene set clustering based on functional annotation and to associate biclusters with specific biological processes. A binary matrix with biclusters along the rows and samples along the columns was generated to perform hierarchical clustering. Samples belonging to respective biclusters were assigned a value of 1. The Hamming distance was used to measure dissimilarity between the biclusters as well as the samples. All tests for enrichments were done using the Fisher's Exact Test. The details of the contingency tables for the tests are provided in Supplementary Methods. The p-values were corrected for multiple testing using the Benjamini–Hochberg false discovery rate method [27]. All enrichments are reported at a significance level of $p < 0.05$ (after correction), unless specified otherwise.

RESULTS

Choosing tuning parameters for TuBA. The choice of TuBA's parameters for a given run determines the final biclusters. Specifically, increasing the size of the percentile set

results in loss of significance for aberrant co-expression signatures found in smaller subsets of samples. This does not imply that it is generally better to choose smaller percentile sets; in fact, a reduction in the size of the percentile set increases the likelihood of matches purely by chance. Thus, the variation of the size of the percentile set from smaller to larger sizes is accompanied by a trade-off between sensitivity (identification of altered transcriptional profiles in small subsets of samples) and specificity (confidence in significance of overlap).

The choice of the second parameter, i.e. the extent of patient/sample overlap, impacts the size of the graph. As we lower the significance of overlap (increasing p-values), new genes and samples get added to the graph, resulting in an increase in the number of edges. Further lowering of the overlap significance results in the addition of more edges to the graph. However, this is accompanied by only a modest gain of new information (especially in terms of samples) in the biclusters themselves. Thus, the choice of cutoff for the significance of overlap is accompanied by a trade-off between the gain of new information in the biclusters and the number of edges that get added to the graph.

For a given choice of the size of percentile set, TuBA generates graphs that visualize the number of added genes, added edges, and added samples as the overlap significance cutoff is varied. Since the experimental platform and the number of samples were different among the analyzed datasets, the values of TuBA's parameters and the total number of biclusters obtained varied (**Fig. 3**). For these choices of the knobs, we obtained 353, 340, and 369 biclusters for the TCGA, METABRIC, and GEO datasets, respectively (**Supplementary Table 1**).

Consistency of TuBA within a dataset. To investigate whether TuBA could consistently discover biclusters within the TCGA RFS cohort, the 908 samples were divided randomly into two groups of 454 samples each five times to generate five pairs

of datasets. TuBA was applied to each dataset pair using a percentile set size of 5% and an overlap significance cut-off $p \leq 10^{-08}$. Pairwise comparisons of biclusters from the five trials showed that that on average 73% biclusters from one dataset in each pair were enriched ($p < 0.001$) in at least one bicluster from the other (**Supplementary Table 2**). We found a significant difference in the number of genes in biclusters that matched among trials, compared to biclusters that did not (Mann-Whitney test $p < 10^{-05}$). Whereas the median size of biclusters that matched was 20 (range: 3–840), the median size of biclusters that did not match was 3 (range: 3–18) (note that 3 is the smallest sized bicluster generated by TuBA). In summary, there was good agreement between biclusters obtained from randomly sampled subsets, exhibiting robust gene co-aberrant signatures.

Consistency of TuBA's among independent datasets. Using common sets of genes, we compared the biclusters obtained from: (i) TCGA and METABRIC, (ii) TCGA and GEO, and (iii) METABRIC and GEO. Pairwise comparisons of biclusters obtained from the two datasets were used to identify the biclusters that shared a significant proportion of their genes ($p < 0.001$). In the TCGA vs. METABRIC comparison, 64% of biclusters obtained in one dataset were enriched in at least one bicluster in the other. In the TCGA vs. GEO comparison, 69% of biclusters obtained in one dataset were enriched in at least one bicluster in the other. Finally, in the METABRIC vs. GEO comparison, 76% of the biclusters obtained in one dataset were enriched in at least one bicluster in the other. Based on the Mann-Whitney test, we found that the biclusters that did not match were significantly smaller than the biclusters that matched between the datasets ($p < 0.001$).

TuBA identifies subtype-specific biclusters. We classified BRCA samples based on the expression levels of the ESR1 (ER) and ERBB2 (HER2) genes into four subtypes: (i) ER-/HER2-, (ii) ER+/HER2-, (iii) ER-/HER2+, and (iv) ER+/HER2+ (where +

corresponds to over expressed and – corresponds to under expressed). A substantial proportion of biclusters were enriched in the ER-/HER2- subtype – 53% for METABRIC (**Fig. 4A and 4B**), 54% for TCGA (**Fig. 4C and Fig. 4D**), and 40% for GEO (**Fig. S6**) (**Supplementary Table 3**).

According to the PAM50 classification, there are five subtypes of BRCA: (i) Basal-like, (ii) Her2-enriched, (iii) Luminal A, (iv) Luminal B, and (v) Normal-like [28]. We observed a significant fraction of biclusters enriched in the basal-like subtype – 52% for METABRIC (**Fig. S4A and S4B**) and 55% for TCGA PAM50 (**Fig. S4C and S4D**). Although tumors of the basal-like or triple negative subtypes accounted for only about 15% of all BRCAs in the population, most of the altered expression profiles captured by our biclusters were in tumors of this subtype.

TuBA identifies down-regulated subtype-specific biclusters in RNA-seq data. RNA sequencing offers a significant advantage over microarray assays. Theoretically, only the depth of sequencing limits the dynamic range of RNA-seq data [29, 30]. Given that TCGA's RNA-seq data has adequate sequencing depth, we expected a reliable quantification of even low expressing transcripts. We therefore applied TuBA to the TCGA datasets to explore transcriptional profiles associated with low expression. We found that 46% biclusters from TCGA were enriched in the ER-/HER2- subtype (**Fig. S5A and S5B**), while 48% biclusters from the TCGA PAM50 dataset were enriched in the basal-like subtype (**Fig. S4E and S4F**). Thus, biclusters associated with low expression were predominantly enriched in the ER-/HER2- or basal-like subtypes. This further underscores the tremendous heterogeneity of altered transcriptional profiles within tumors of this subtype.

TuBA highlights biclusters with proximally located genes. We observed that several biclusters discovered by TuBA across the three datasets comprise genes that are

proximally located on the chromosomes, suggesting copy number amplification (CNA) as an underlying mechanism. Copy number data was used to calculate the significance of the proportion of samples present in each bicluster that exhibited copy number gains. For each gene in a given bicluster, we computed a p-value for the significance of the proportion of samples with CNA present in the bicluster. These p-values were then combined using Fisher's method to yield a single p-value for the bicluster. This showed that 56% and 64% of biclusters from the METABRIC and TCGA datasets respectively were enriched for CNA. Closer scrutiny revealed that only 60 (18%) biclusters from METABRIC were associated with CNA of proximally located genes (Fig. 4A), the remaining biclusters associated with CNA were enriched in genes from distant chromosomal locations (Fig. 4B). Similarly, 112 (32%) biclusters from TCGA were associated with CNA of proximally located genes (Fig. 4C). Many of these biclusters were associated with loci previously identified to exhibit copy number gains in BRCA [31, 32]. In order to explore the association between the biclusters obtained from the low expression analysis and loss of copy number, we repeated the copy number analysis described above. We observed that 52% biclusters from the TCGA dataset were enriched in copy number losses. However, only 21 biclusters contained genes located on the same chromosome (Fig. S5A), the remaining biclusters associated with copy number loss were enriched in genes from distant chromosomal locations. Similar analyses for PAM50 subtype enrichment for METABRIC and TCGA are summarized in Fig. S4.

To compare CNA associated biclusters between TCGA and METABRIC, we prepared two datasets that contained genes that were common in the two cohorts (17,209 genes). Pairwise comparison of the set of genes in the CNA enriched biclusters between the two datasets revealed that ~61% of biclusters from TCGA matched at least one CNA

associated bicluster from METABRIC. On the other hand, 91% of biclusters from METABRIC were enriched in at least one CNA associated bicluster from TCGA. This suggests that most of the CNA enriched biclusters identified in the METABRIC microarray dataset were independently identified in the RNASeq dataset of TCGA.

We also observed some biclusters with proximally located genes that were not associated with gain in copy number. For TCGA, 14 biclusters out of 353 consisted of genes located proximally, while 18 biclusters out of 340 for METABRIC consisted of genes located near each other. Details of the genes and subtype-specific enrichments for some of these biclusters are summarized in Supplementary Table 4. Examples of biclusters from this category include the biclusters consisting of genes from the Cancer-Testis antigens family – MAGEA2, MAGEA3, MAGEA6, MAGEA10, CSAG1, CSAG2, CSAG3 (Xq28)/CT45A3, CT45A5, CT45A6 (Xq26.3). These genes are known to be aberrantly expressed in triple negative breast tumors [33] as well as in some other tumor types [34].

TuBA identifies biclusters associated with non-tumor expression signatures. We also discovered biclusters that appeared to be associated with non-tumor cells. For instance, biclusters associated with immune response were among the largest identified independently in all three datasets. The top five Gene Ontology – Biological Processes (GO-BP) terms for the bicluster associated with immune response were: T cell costimulation, T cell receptor signaling pathway, T cell activation, regulation of immune response, and positive regulation of T cell proliferation (**Supplementary Table 5**). This indicates immune cell infiltration in a significant number of tumor samples. To corroborate this, we stratified TCGA samples based on their ESTIMATE [35] scores for the infiltration level of immune cells in tumor tissues into three groups – (i) top 25 percentile, (ii) intermediate 50 percentile, and (iii) bottom 25 percentile – and verified that

these biclusters associated with immune response were enriched with samples with the highest levels of immune infiltration ($p < 0.001$)

For all three datasets, we also observed a bicluster associated with the stromal adipose tissue. The top 5 GO-BP terms for this bicluster were: response to glucose, triglyceride biosynthetic process, triglyceride catabolic process, retinoid metabolic process, and retinol metabolic process. An analysis based on the ESTIMATE scores for the level of stromal cells present in tumor tissue of TCGA samples confirmed that this bicluster was enriched within the top 25 percentile samples for stromal cell level. Subtype enrichment revealed that the bicluster was enriched in ER-/HER2-, basal-like (PAM50), and normal-like (PAM50) subtypes.

TuBA's proximity measure was applied to gene expression data from 214 normal breast tissue samples from the Genotype-Tissue Expression (GTEx) public dataset. We observed that only 6.75% of biclusters obtained for the TCGA versus GTEx comparison were enriched in gene-pair associations identified in the GTEx dataset. The bicluster associated with the adipose tissue signature was one of the biclusters found enriched in GTEx. Another group of biclusters enriched in the three cancer datasets as well as in GTEx, were those associated with translation and ribosomal assembly. The top 5 GO-BP terms for these biclusters were: translation, rRNA processing, ribosomal small subunit biogenesis, ribosomal large subunit assembly, and ribosomal large subunit biogenesis. These biclusters were enriched in the ER-/HER2- subtype ($p < 0.001$).

TuBA identifies clinically associated biclusters. We performed a Kaplan-Meier (KM) analysis of recurrence free survival (RFS), comparing the patients present in each bicluster to the rest for METABRIC and GEO. (The number of patients with incidence of recurrence in TCGA was insufficient for this kind of survival analysis to be statistically robust.) As expected for METABRIC, patients in the bicluster associated with the HER2

amplicon (17q12) had significantly shorter RFS time compared to the rest (**Fig. S8**). This is partly because patients in the METABRIC study were enrolled before the general availability of trastuzumab [36].

We also observed biclusters associated with copy number gains at the 8q24.3 locus in all three datasets. These patients also had significantly shorter RFS times compared to those patients whose tumors did not have amplification of this locus (**Fig. 5A**, **5B** and **5C**). A similar result was obtained when we restricted the samples to ER+/HER2- tumors, validating an earlier observation that copy number gain of the 8q24.3 locus may confer resistance to ER targeted therapy [37]. We note, however, that biclusters with amplification of the 8q24.3 locus were enriched in the ER-/HER2- subtype ($p < 0.001$). Hence, amplification of this locus may be even more relevant in determining treatment for patients with ER-/HER2- breast cancers assigned into an intermediate (ambiguous) risk class by Oncotype DX [37]. Genes at 8q24.3 that may be considered promising candidates based on their degrees in the biclusters include *PUF60*, *EXOSC4*, *COMMD5*, and *HSF1*. Specifically, *PUF60* is an RNA-binding protein known to contribute to tumor progression by enabling increased *MYC* expression and greater resistance to apoptosis [38].

For both METABRIC and GEO, patients in biclusters associated with copy number gains of the 8p11.21-p11.23 loci had significantly shorter RFS times compared to patients without amplification of this locus (**Fig. 5D**, **5E** and **5F**). We found that patients in this bicluster were enriched in the Luminal B subtype, which has poorer prognosis than the Luminal A subtype among ER+/HER2- tumors [39]. This suggested that amplification of the 8p11.21-p11.23 loci might be another marker of potential failure of ER targeted therapy.

Similarly, we found that patients whose tumors have copy number gains of the 17q22-q23.3 locus had significantly shorter RFS times compared to patients whose tumors do not exhibit such a copy number gain (**Fig. 5G**, **5H** and **5I**). For METABRIC, this cohort was enriched in the Luminal B (PAM50), ER+/HER2+, and ER-/HER2+ subtypes ($p < 0.001$). For GEO, this cohort was enriched in the ER+/HER2+ and ER-/HER2+ subtypes ($p < 0.05$). This suggests that amplification of this locus may confer additional risk of recurrence in HER2+ breast cancers.

Note that the biclusters discussed above were not the only ones that exhibited differential relapse outcomes. For METABRIC, 61 biclusters out of 340 were found to exhibit differential relapse outcomes for the patients present in the biclusters. Out of these 61 biclusters, 69% were enriched in the ER-/HER2- subtype (64% for basal-like) with a significant proportion (67%) of these associated with copy number gains. For GEO, there were 48 such biclusters (13%) that exhibited differential relapse outcomes, 25% of these were enriched in the ER-/HER2- subtype.

Tests for enrichment of biclusters in tumors of higher grades revealed that eight biclusters from TCGA were enriched in tumors of grade 3C. Some of these biclusters were associated with GO-BP terms related to angiogenesis, vasculogenesis, blood vessel maturation etc. For METABRIC, four biclusters were enriched in tumors of grade 3, out of which two were associated with the HER2 amplicon (17q12). For GEO, 68 biclusters were enriched in tumors of grade 3, including biclusters associated with copy number gains at the HER2 amplicon.

We also looked at the lymph node status of patients and observed that four biclusters in TCGA were enriched in samples with positive lymph node status in the corresponding patients. One was associated with the HER2 amplicon, while the others were associated with copy number gains at the 8q22.1-q22.3 loci, 17q23.1-q23.3 loci and the 19q13.43

locus, respectively. Similarly in METABRIC, we observed four biclusters enriched in samples with positive lymph node status in the corresponding patients. Two of them were associated with copy number gains at the HER2 amplicon. The others were associated with copy number gains at 19q13.11-q13.12 and 1q21.3-q25.1, respectively. Interestingly, biclusters associated with copy number gains at 8q24.3, 8p11.21-p11.23, and 17q22-q23.3 that exhibited poor RFS outcomes were not enriched in tumors of higher grades or in patients with positive lymph node status in any of the three datasets. In case of METABRIC, we additionally confirmed that none of these biclusters (8q24.3, 8p11.21-p11.23, 17q23.1-q23.3) were among the 36 biclusters enriched in samples with the poorest expected 5-year survival outcome (Nottingham Prognostic Index (NPI): > 5.4) [40, 41]. This highlights the importance of these altered transcriptomic signatures for reclassification of patients into the category with higher risk of recurrence.

Hierarchical clustering of biclusters reveals shared mechanisms. Sample membership based hierarchical clustering of biclusters revealed distinct groups of biclusters that presumably share common functional mechanisms (**Fig. 6**). These included clusters associated with cell cycle and proliferation, immune response, cell adhesion (extracellular matrix), translation, mitochondrial translation, and ribosomal RNA processing pathways. Since a significant fraction of our biclusters were associated with copy number alterations, we also found distinct groups of biclusters associated with significant copy number changes such as the ones associated with the HER2 amplicon, the 8p11.21-p11.23 loci, or the 8q24.3 locus.

Similarly, we used hierarchical clustering to group samples that were enriched in similar sets of biclusters, highlighting differential clinical outcomes. In particular, we observed two sets of samples enriched in biclusters associated with copy number gains at the 8q24.3 locus. In one group, the samples were enriched in biclusters related to immune

response; this group showed significantly lower incidence of recurrence compared to those without enrichment in immune response-related biclusters. Both of these sets of samples were enriched in biclusters associated with cell division and proliferation. In contrast, we observed a cluster of samples enriched in biclusters associated with 8q24.3 copy number gain and a number of other loci, yet not enriched in biclusters associated with cell division and proliferation; this group exhibited low incidence of recurrence. We also observed a cluster of samples with significantly poor RFS that were enriched in biclusters associated with copy number gains at 17q25.1-q25.3 and in biclusters associated with cell division and proliferation.

DISCUSSION

Global clustering approaches have successfully unveiled distinct disease subtypes in tumors, prompting the community to look beyond traditional clinico-pathological signatures to identify relevant disease processes. However, the extensive heterogeneity, even within tumors of a given subtype, confounds the identification of many altered transcriptional programs by such unsupervised clustering methods. While there are many biclustering approaches that attempt to address this limitation, most of them are not customized to identify the heterogeneous transcriptional profiles that are expressed aberrantly by tumors. In this paper, we introduce an algorithm called TuBA based on a proximity measure specifically designed to extract gene co-expression signatures that correspond to the extremes of expression (both high and low for RNASeq data, and high for any other platform), thereby preferentially identifying gene co-aberrant signatures associated with the disease states of tumors. The identification of altered transcriptional profiles can be particularly relevant for those tumors that have so far eluded targeted drug development for therapy. This is exemplified by tumors of the basal-like or triple negative subtypes for BRCA. Although these tumors account for only ~15% of all BRCAs

in the population, a significant fraction of biclusters identified by TuBA corresponded to alterations associated with tumors of these subtypes. For each dataset, a simple estimation of enrichment of samples in a given bicluster within any other bicluster, revealed that the samples in the biclusters corresponding to CNA at 8p11.21-p11.23 or 17q12 were enriched ($p < 0.001$) independently in ~5% of all biclusters for both TCGA and METABRIC, respectively. In sharp contrast, 30–40% of all biclusters were enriched in samples with copy number gains at the 8q24.3 locsus ($p < 0.001$). Additionally, 51% of all biclusters obtained from the low expression analysis of TCGA were enriched in the samples corresponding to the 8q24.3 bicluster. Previous studies have also identified the amplicon at 8q24.3 by Representational Difference Analysis as a location of oncogenic alterations in breast cancer that can occur independent of neighboring *MYC* amplifications [42]. Although the 8q24.3 bicluster itself is enriched with ER-/HER2- samples, these observations together with poor RFS outcome observed independently in both METABRIC and GEO highlight this locus as a good prognostic marker for BRCA tumors, irrespective of subtype.

Apart from highlighting the heterogeneity of CNA-associated alterations in tumors of the ER-/HER2- subtype or basal-like subtype, TuBA offered a glimpse into the utility, the limitations, and the potential pitfalls with the current subtype classification approaches. In the ER/HER2-based subtype enrichment, we observed that a significant proportion of biclusters were not specifically enriched in any one of the four subtypes. For instance, several CNA-associated biclusters from chromosome eight were not subtype-enriched. In the case of PAM50 subtype classification however, we observed that most of these biclusters were enriched in the Luminal B subtype for METABRIC (and to a limited extent for TCGA). While this may appear to indicate that PAM50 does offer an improvement on the traditional clinico-pathological approach to subtype classification, it

however fails to classify several samples associated with overexpression of ERBB2 as HER2-positive. As a consequence, several of our biclusters associated with the HER2 amplicon and copy number gains in the neighboring locations on chromosome 17 (17q.21.1-q21.2 and 17q21.32-q21.33), for both METABRIC and TCGA, were observed to be enriched in the Luminal B subtype. This corroborates the modest agreement reported in [28], as well as disagreements in later studies [43]. Given that trastuzumab is a clinically proven therapeutic drug for HER2+ tumors, misclassification of these patients into any other subtype can be highly disadvantageous.

Change in copy number is often not a sufficient condition for elevated (or suppressed) expression levels of transcripts, as there are multiple layers of regulation of transcription in cells [44, 45]. TuBA specifically identifies sets of genes with copy number changes that are transcriptionally active (or inactive), filtering out the ones that are unlikely to influence disease progression. Moreover, the graph-based approach allows us to infer the relative importance of each gene within a bicluster, based on its degree. In the case of high expression analysis, the degree of each gene is an indicator of how frequently it is expressed aberrantly by the subset of samples that comprise a bicluster. As an example, consider the CNA-associated bicluster from TCGA corresponding to gains at the 8q22.1-q22.3 loci. The bicluster exhibited enrichment in lymph node positive patients (the corresponding bicluster in METABRIC has a significance level of $p = 0.052$ for patients with positive lymph node status). The gene with the highest degree in the bicluster was *MTDH* (metadherin), which has been shown to be associated with increased chemoresistance and metastasis in BRCA [46-48].

Clustering analysis of biclusters and samples based on the membership of samples within biclusters allowed us to identify the sites that were altered concomitantly within the same subsets of samples. Moreover, we improved our perspective on the tumor

microenvironment in the subsets of samples that exhibit non-tumor associated signatures (such as immune, extracellular matrix, etc.). Differences in disease progression due to distinct microenvironments in tumors with similar transcriptional alterations can help us better understand the potential role of the microenvironment within the context of tumors harboring these specific alterations. For instance, we noticed a difference in RFS outcomes between two groups of patients that exhibit copy number gains at 8q24.3. The group that was additionally associated with an immune response signature was observed to have better RFS outcomes compared to the group that did not exhibit a strong association with the immune response.

A limitation of TuBA is that it can only be applied reliably for relatively large datasets. Depending on cohort heterogeneity, some of the overlaps between percentile sets may not be significant in smaller datasets (**Supplementary Methods**). However, the deliberate design of our proximity measure leveraging the size of the datasets offers a significant benefit – it not only enables the identification of the plethora of gene co-aberrations associated with the tumors, but also allows us to estimate the extent or prevalence of the heterogeneity and various altered signatures in the population. This is where the tunable aspect of TuBA becomes relevant – the two knobs should be viewed as valuable aids that help estimate the extents of the prevalence of various alterations in the tumor population and their clinical relevance. Although transcriptomic changes are not the ultimate determinants of progression, our algorithm holds the promise to improve therapeutic selection and design by identifying significantly altered transcriptional patterns associated with tumors.

REFERENCES

1. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998;95(25):14863-8. Epub 1998/12/09. PubMed PMID: 9843981; PubMed Central PMCID: PMCPMC24541.

2. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403(6769):503-11. Epub 2000/02/17. doi: 10.1038/35000501. PubMed PMID: 10676951.

3. Roth FP, Hughes JD, Estep PW, Church GM. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol. 1998;16(10):939-45. Epub 1998/10/27. doi: 10.1038/nbt1098-939. PubMed PMID: 9788350.

4. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531-7. Epub 1999/10/16. PubMed PMID: 10521349.

5. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci U S A. 1999;96(16):9212-7. Epub 1999/08/04. PubMed PMID: 10430922; PubMed Central PMCID: PMCPMC17759.

6. Cheng Y, Church GM. Biclustering of expression data. Proc Int Conf Intell Syst Mol Biol. 2000;8:93-103. Epub 2000/09/08. PubMed PMID: 10977070.

7. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinform. 2004;1(1):24-45. Epub 2006/10/20. doi: 10.1109/TCBB.2004.2. PubMed PMID: 17048406.

8. Eren K, Deveci M, Kucuktunc O, Catalyurek UV. A comparative analysis of biclustering algorithms for gene expression data. Brief Bioinform. 2013;14(3):279-92. Epub 2012/07/10. doi: 10.1093/bib/bbs032. PubMed PMID: 22772837; PubMed Central PMCID: PMCPMC3659300.

9. Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. Bioinformatics. 2002;18 Suppl 1:S136-44. Epub 2002/08/10. PubMed PMID: 12169541.

10. Oghabian A, Kilpinen S, Hautaniemi S, Czeizler E. Biclustering methods: biological relevance and application in gene expression analysis. PLoS One. 2014;9(3):e90801. Epub 2014/03/22. doi: 10.1371/journal.pone.0090801. PubMed PMID: 24651574; PubMed Central PMCID: PMCPMC3961251.

11. Yoon S, Nardini C, Benini L, De Micheli G. Discovering coherent biclusters from gene expression data using zero-suppressed binary decision diagrams. IEEE/ACM Trans Comput Biol Bioinform. 2005;2(4):339-54. Epub 2006/10/19. doi: 10.1109/TCBB.2005.55. PubMed PMID: 17044171.

12. Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci U S A. 2000;97(22):12079-84. Epub 2000/10/18. doi: 10.1073/pnas.210134797. PubMed PMID: 11035779; PubMed Central PMCID: PMCPMC17297.

13. Sheng Q, Moreau Y, De Moor B. Biclustering microarray data by Gibbs sampling. Bioinformatics. 2003;19 Suppl 2:ii196-205. Epub 2003/10/10. PubMed PMID: 14534190.

14. Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, Kasim A, et al. FABIA: factor analysis for bicluster acquisition. Bioinformatics. 2010;26(12):1520-7. Epub 2010/04/27. doi: 10.1093/bioinformatics/btq227. PubMed PMID: 20418340; PubMed Central PMCID: PMCPMC2881408.

15. Bron C, Kerbosch J. Algorithm 457: finding all cliques of an undirected graph. Communications of the ACM. 1973;16(9):575-7. doi: 10.1145/362342.362367.

16. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61-70. Epub 2012/09/25. doi: 10.1038/nature11412. PubMed PMID: 23000897; PubMed Central PMCID: PMCPMC3465532.

17. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41. Epub 2011/04/30. doi: 10.1186/gb-2011-12-4-r41. PubMed PMID: 21527027; PubMed Central PMCID: PMCPMC3218867.

18. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401-4. Epub 2012/05/17. doi: 10.1158/2159-8290.CD-12-0095. PubMed PMID: 22588877; PubMed Central PMCID: PMCPMC3956037.

19. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1. Epub 2013/04/04. doi: 10.1126/scisignal.2004088. PubMed PMID: 23550210; PubMed Central PMCID: PMCPMC4160307.

20. Gyorffy B, Schafer R. Meta-analysis of gene expression profiles related to relapse-free survival in 1,079 breast cancer patients. Breast Cancer Res Treat. 2009;118(3):433-41. Epub 2008/12/05. doi: 10.1007/s10549-008-0242-8. PubMed PMID: 19052860.

21. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106. Epub 2010/10/29. doi: 10.1186/gb-2010-11-10-r106. PubMed PMID: 20979621; PubMed Central PMCID: PMCPMC3218662.

22. R Core Team. R: A Language and Environment for Statistical Computing. 2016.

23. Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. InterJournal. 2006;Complex Systems:1695.

24. Wickham H. The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software. 2011;40(1):1-29.

25. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. 2009.

26. Subhash S, Kanduri C. GeneSCF: a real-time based functional enrichment tool with support for multiple organisms. BMC Bioinformatics. 2016;17(1):365. Epub 2016/09/14. doi: 10.1186/s12859-016-1250-z. PubMed PMID: 27618934; PubMed Central PMCID: PMCPMC5020511.

27. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 1995;57(1):289-300.

28. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160-7. Epub 2009/02/11. doi: 10.1200/JCO.2008.18.1370. PubMed PMID: 19204204; PubMed Central PMCID: PMCPMC2667820.

29. Marguerat S, Bahler J. RNA-seq: from technology to biology. Cell Mol Life Sci. 2010;67(4):569-79. Epub 2009/10/28. doi: 10.1007/s00018-009-0180-6. PubMed PMID: 19859660; PubMed Central PMCID: PMCPMC2809939.

30. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621-8. Epub 2008/06/03. doi: 10.1038/nmeth.1226. PubMed PMID: 18516045.

31. Kallioniemi A, Kallioniemi OP, Piper J, Tanner M, Stokke T, Chen L, et al. Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. Proc Natl Acad Sci U S A. 1994;91(6):2156-60. Epub 1994/03/15. PubMed PMID: 8134364; PubMed Central PMCID: PMCPMC43329.

32. Kao J, Salari K, Bocanegra M, Choi YL, Girard L, Gandhi J, et al. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. PLoS One. 2009;4(7):e6146. Epub 2009/07/08. doi: 10.1371/journal.pone.0006146. PubMed PMID: 19582160; PubMed Central PMCID: PMCPMC2702084.

33. Curigliano G, Viale G, Ghioni M, Jungbluth AA, Bagnardi V, Spagnoli GC, et al. Cancer-testis antigen expression in triple-negative breast cancer. Ann Oncol. 2011;22(1):98-103. Epub 2010/07/09. doi: 10.1093/annonc/mdq325. PubMed PMID: 20610479.

34. Simpson AJ, Caballero OL, Jungbluth A, Chen YT, Old LJ. Cancer/testis antigens, gametogenesis and cancer. Nat Rev Cancer. 2005;5(8):615-25. Epub 2005/07/22. doi: 10.1038/nrc1669. PubMed PMID: 16034368.

35. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612. Epub 2013/10/12. doi: 10.1038/ncomms3612. PubMed PMID: 24113773; PubMed Central PMCID: PMCPMC3826632.

36. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346-52. Epub 2012/04/24. doi:

10.1038/nature10983. PubMed PMID: 22522925; PubMed Central PMCID: PMCPMC3440846.

37. Bilal E, Vassallo K, Toppmeyer D, Barnard N, Rye IH, Almendro V, et al. Amplified loci on chromosomes 8 and 17 predict early relapse in ER-positive breast cancers. PLoS One. 2012;7(6):e38575. Epub 2012/06/22. doi: 10.1371/journal.pone.0038575. PubMed PMID: 22719901; PubMed Central PMCID: PMCPMC3374812.

38. Wang J, Liu Q, Shyr Y. Dysregulated transcription across diverse cancer types reveals the importance of RNA-binding protein in carcinogenesis. BMC Genomics. 2015;16 Suppl 7:S5. Epub 2015/06/24. doi: 10.1186/1471-2164-16-S7-S5. PubMed PMID: 26100984; PubMed Central PMCID: PMCPMC4474540.

39. Inic Z, Zegarac M, Inic M, Markovic I, Kozomara Z, Djurisic I, et al. Difference between Luminal A and Luminal B Subtypes According to Ki-67, Tumor Size, and Progesterone Receptor Negativity Providing Prognostic Information. Clin Med Insights Oncol. 2014;8:107-11. Epub 2014/09/25. doi: 10.4137/CMO.S18006. PubMed PMID: 25249766; PubMed Central PMCID: PMCPMC4167319.

40. Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, et al. A prognostic index in primary breast cancer. Br J Cancer. 1982;45(3):361-6. Epub 1982/03/01. PubMed PMID: 7073932; PubMed Central PMCID: PMCPMC2010939.

41. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. Breast Cancer Res Treat. 1992;22(3):207-19. Epub 1992/01/01. PubMed PMID: 1391987.

42. Mu D, Chen L, Zhang X, See LH, Koch CM, Yen C, et al. Genomic amplification and oncogenic properties of the KCNK9 potassium channel gene. Cancer Cell. 2003;3(3):297-302. Epub 2003/04/05. PubMed PMID: 12676587.

43. Guiu S, Michiels S, Andre F, Cortes J, Denkert C, Di Leo A, et al. Molecular subclasses of breast cancer: how do we define them? The IMPAKT 2012 Working Group Statement. Ann Oncol. 2012;23(12):2997-3006. Epub 2012/11/21. doi: 10.1093/annonc/mds586. PubMed PMID: 23166150.

44. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013;152(6):1237-51. Epub 2013/03/19. doi: 10.1016/j.cell.2013.02.014. PubMed PMID: 23498934; PubMed Central PMCID: PMCPMC3640494.

45. Lelli KM, Slattery M, Mann RS. Disentangling the many layers of eukaryotic transcriptional regulation. Annu Rev Genet. 2012;46:43-68. Epub 2012/09/01. doi: 10.1146/annurev-genet-110711-155437. PubMed PMID: 22934649; PubMed Central PMCID: PMCPMC4295906.

46. Wan L, Kang Y. Pleiotropic roles of AEG-1/MTDH/LYRIC in breast cancer. Adv Cancer Res. 2013;120:113-34. Epub 2013/07/31. doi: 10.1016/B978-0-12-401676-7.00004-8. PubMed PMID: 23889989.

47. Song Z, Wang Y, Li C, Zhang D, Wang X. Molecular Modification of Metadherin/MTDH Impacts the Sensitivity of Breast Cancer to Doxorubicin. PLoS One. 2015;10(5):e0127599. Epub 2015/05/21. doi: 10.1371/journal.pone.0127599. PubMed PMID: 25993398; PubMed Central PMCID: PMCPMC4437901.

48. Shi X, Wang X. The role of MTDH/AEG-1 in the progression of cancer. Int J Clin Exp Med. 2015;8(4):4795-807. Epub 2015/07/02. PubMed PMID: 26131054; PubMed Central PMCID: PMCPMC4484038.

FIGURE LEGENDS

Fig. 1. Schematic representation of TuBA's proximity measure. For each gene, samples are arranged in increasing order of expression levels and those corresponding to a fixed percentile set (top or bottom) are compared between each pair of genes as shown. The gene pairs that share a significant number of samples are represented as nodes linked by edges, which represent the samples.

Fig. 2. TuBA's schematics. (A) Flowchart of the pipeline for TuBA. (B) Schematic representation of the graph-based approach to discover biclusters.

Fig. 3. The effect of TuBA's parameters on the number of genes and samples in biclusters. Plots for the number of genes added to all biclusters for every incremental decrease in the significance level for overlap ($-\log_{10}(p)$), the number of samples in graph at different significance levels of overlap and the total number of edges in graph at different significance levels of overlap corresponding to a percentile set size of 5% for (A) METABRIC, (B) TCGA, and (C) GEO datasets, respectively.

Fig. 4. Subtype enrichment of CNV-associated and non-CNV biclusters. Enrichment of biclusters consisting of proximally located genes with copy number gains in the four subtypes based on ER/HER2 status for (A) METABRIC and (C) TCGA, respectively. The biclusters are represented by horizontal bars in each panel, color-coded according to the chromosome number of their constituent genes. Panels (B) and (D) show the remaining biclusters arranged according to their serial numbers in Supplementary Table 3 for METABRIC and TCGA, respectively. The ones that are associated with copy number (CN) gains of genes located at distant chromosomal sites are shown in red, while the rest are shown in black. Note, the thickness of the bar in each figure depends on the total number of biclusters displayed in that figure and so does not represent its chromosomal extent.

Fig. 5. Clinically associated biclusters. Kaplan-Meier survival curves for the set of patients in the bicluster (red) compared to the remaining set of patients (blue) for METABRIC and GEO datasets, together with the graphs corresponding to the biclusters for 8q24.3 (A, B, and C), 8p11.22-p11.23 (D, E, and F) and 17q22-q23.3 (G, H, and I).

Fig. 6. Hierarchical clustering of biclusters-samples. Hamming distance was applied to the biclusters-samples binary matrix for (A) TCGA and (B) METABRIC RFS datasets, respectively. The clusters of samples marked by green, brown, and cyan on top in panel (B) exhibit poor recurrence free survival. The green and brown clusters are associated with copy number gains at 8q24.3, while the cyan cluster is associated with copy number gains at 17q25.1-q25.3. Additionally, all three of them were enriched in gene signatures associated with cellular division and proliferation.

Figure 1

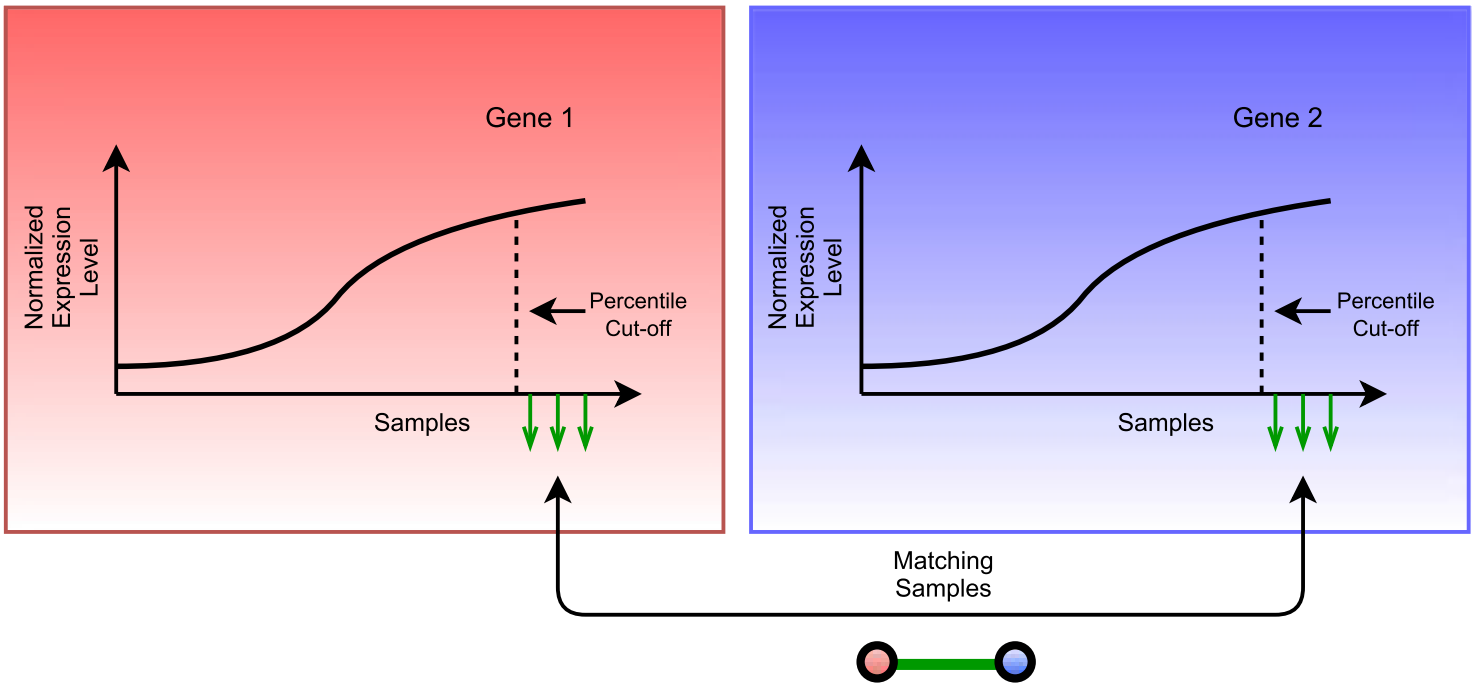

Gene 1
Normalized Expression Level
Percentile Cut-off
Samples
Gene 2
Normalized Expression Level
Percentile Cut-off
Samples
Matching Samples

Figure 2

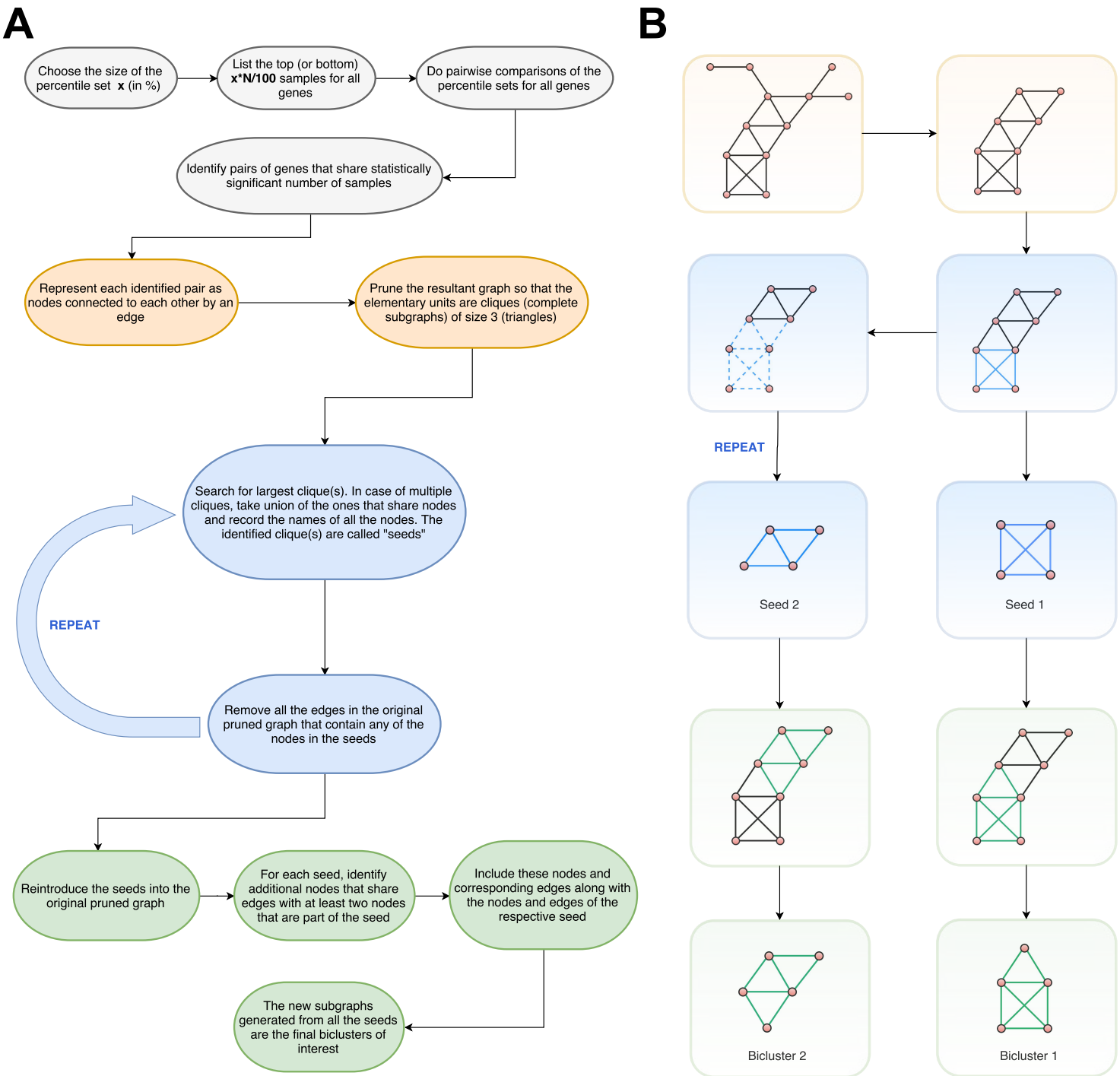

A
Choose the size of the percentile set x (in %)
List the top (or bottom) x*N/100 samples for all genes
Do pairwise comparisons of the percentile sets for all genes
Identify pairs of genes that share statistically significant number of samples
Represent each identified pair as nodes connected to each other by an edge
Prune the resultant graph so that the elementary units are cliques (complete subgraphs) of size 3 (triangles)
Search for largest clique(s). In case of multiple cliques, take union of the ones that share nodes and record the names of all the nodes. The identified clique(s) are called "seeds"
REPEAT
Remove all the edges in the original pruned graph that contain any of the nodes in the seeds
Reintroduce the seeds into the original pruned graph
For each seed, identify additional nodes that share edges with at least two nodes that are part of the seed
Include these nodes and corresponding edges along with the nodes and edges of the respective seed
The new subgraphs generated from all the seeds are the final biclusters of interest
B
REPEAT
Seed 2
Seed 1
Bicluster 2
Bicluster 1

Figure 3

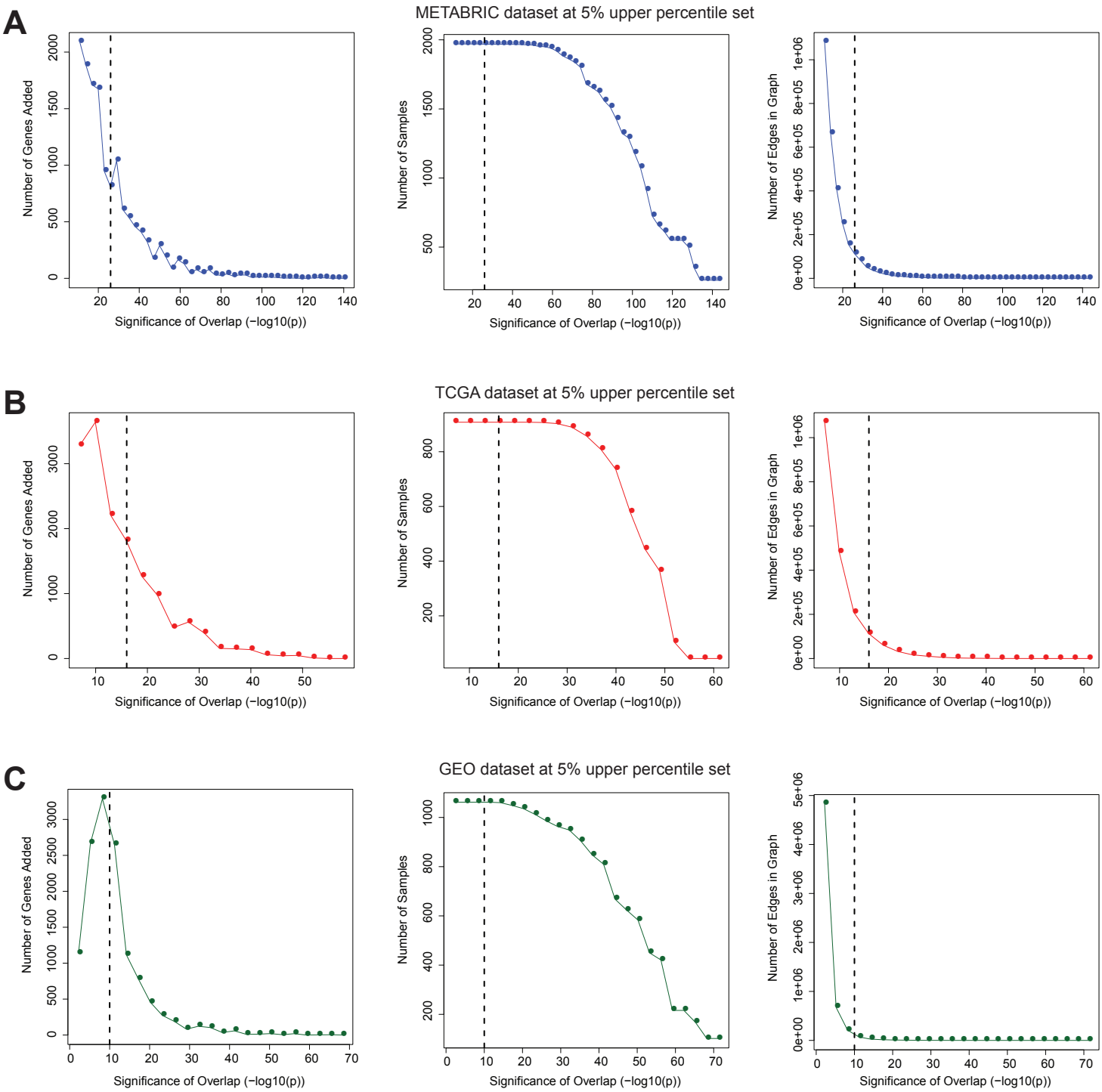

A
METABRIC dataset at 5% upper percentile set
Number of Genes Added
Number of Samples
Number of Edges in Graph
Significance of Overlap (−log10(p))
B
TCGA dataset at 5% upper percentile set
Number of Genes Added
Number of Samples
Number of Edges in Graph
Significance of Overlap (−log10(p))
C
GEO dataset at 5% upper percentile set
Number of Genes Added
Number of Samples
Number of Edges in Graph
Significance of Overlap (−log10(p))

Figure 4

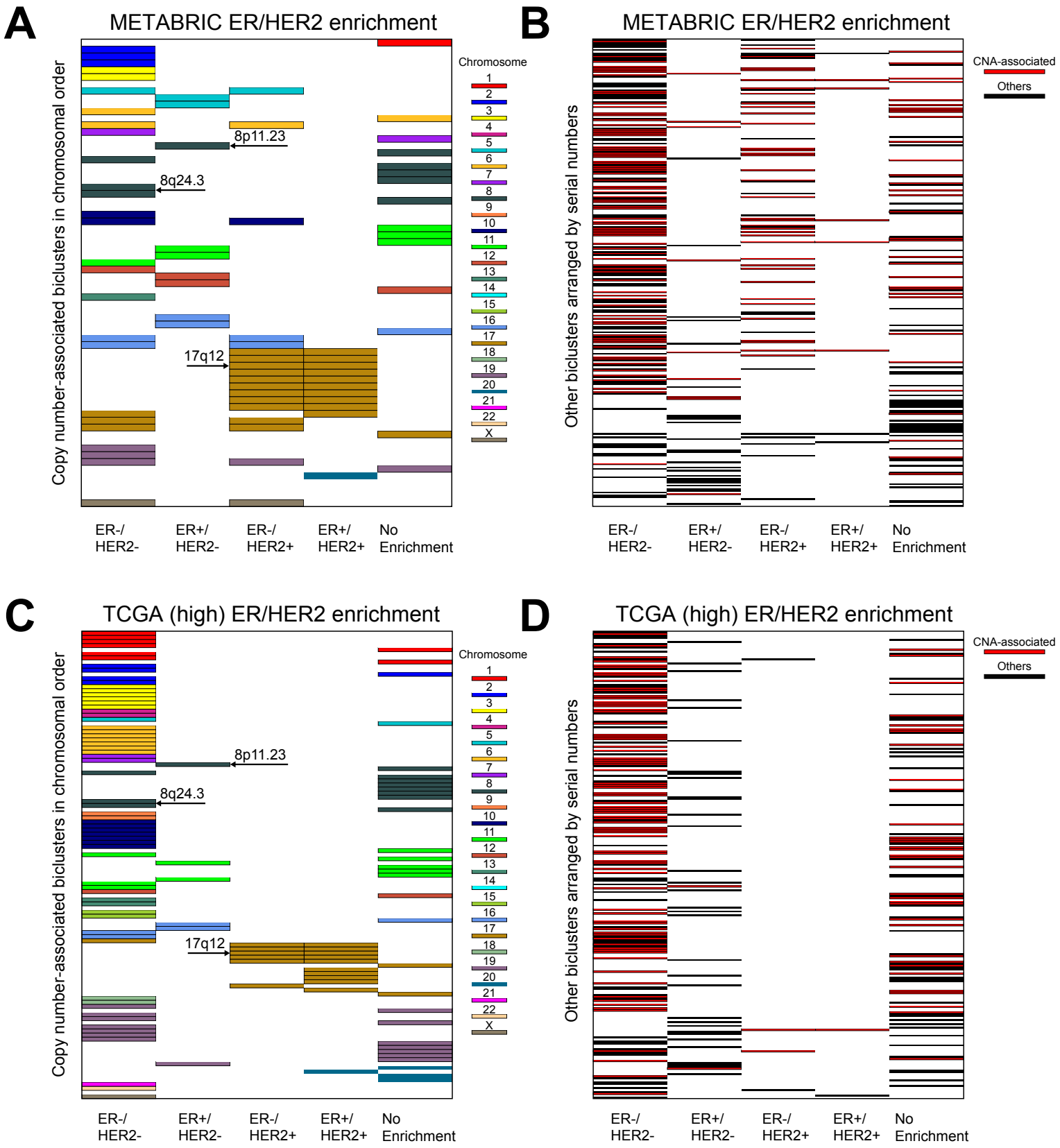

A
METABRIC ER/HER2 enrichment
Copy number-associated biclusters in chromosomal order
8p11.23
8q24.3
17q12
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
ER-/ HER2-
ER+/ HER2-
ER-/ HER2+
ER+/ HER2+
No Enrichment
B
METABRIC ER/HER2 enrichment
Other biclusters arranged by serial numbers
CNA-associated
Others
ER-/ HER2-
ER+/ HER2-
ER-/ HER2+
ER+/ HER2+
No Enrichment
C
TCGA (high) ER/HER2 enrichment
Copy number-associated biclusters in chromosomal order
8p11.23
8q24.3
17q12
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
ER-/ HER2-
ER+/ HER2-
ER-/ HER2+
ER+/ HER2+
No Enrichment
D
TCGA (high) ER/HER2 enrichment
Other biclusters arranged by serial numbers
CNA-associated
Others
ER-/ HER2-
ER+/ HER2-
ER-/ HER2+
ER+/ HER2+
No Enrichment

Figure 5

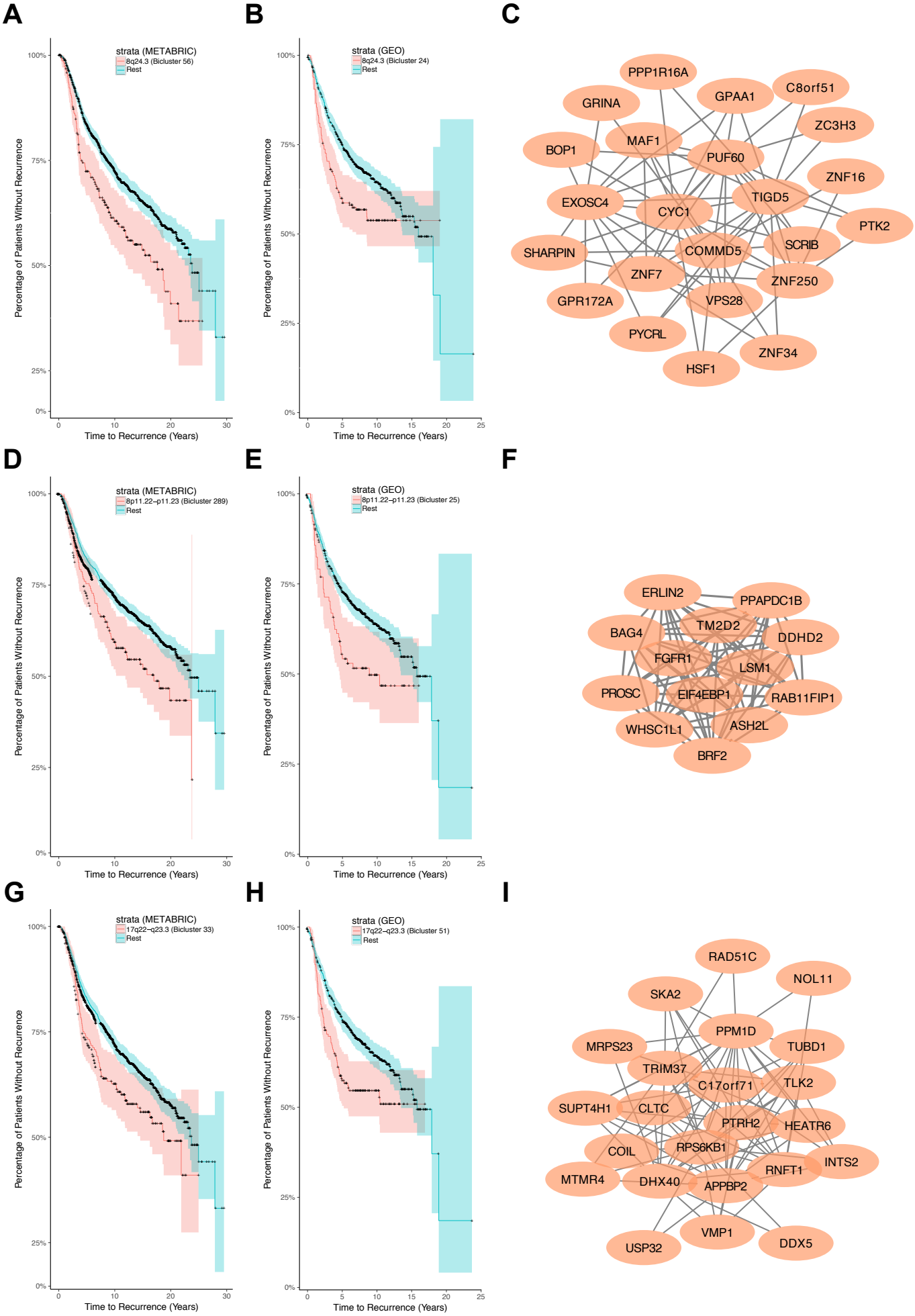
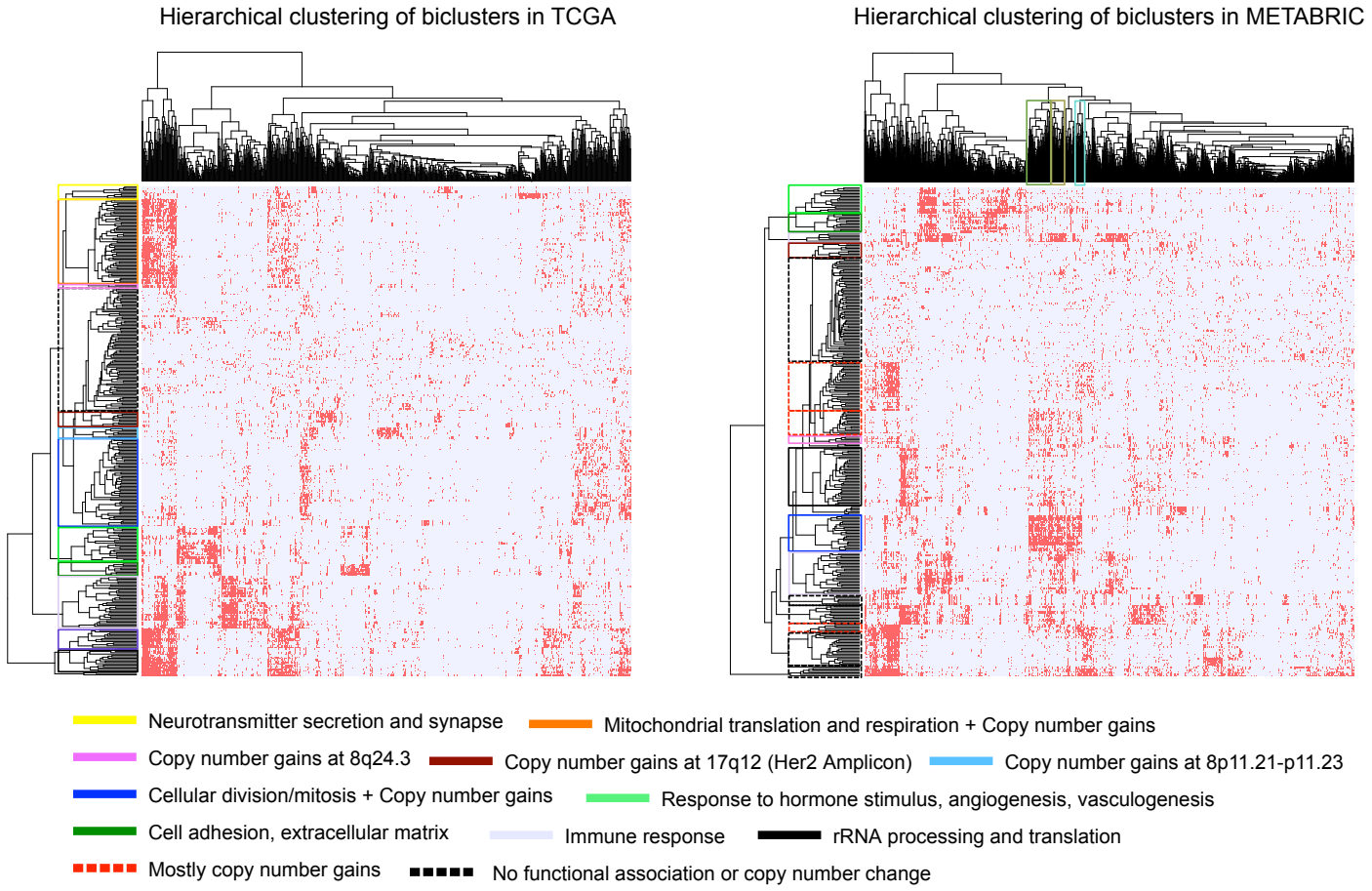

A
strata (METABRIC)
8q24.3 (Bicluster 56)
Rest
Percentage of Patients Without Recurrence
Time to Recurrence (Years)
B
strata (GEO)
8q24.3 (Bicluster 24)
Rest
Percentage of Patients Without Recurrence
Time to Recurrence (Years)
C
PPP1R16A
GRINA
GPAA1
C8orf51
ZC3H3
BOP1
MAF1
PUF60
ZNF16
EXOSC4
CYC1
TIGD5
PTK2
SHARPIN
COMMD5
SCRIB
ZNF7
ZNF250
GPR172A
VPS28
PYCRL
ZNF34
HSF1
D
strata (METABRIC)
8p11.22–p11.23 (Bicluster 289)
Rest
Percentage of Patients Without Recurrence
Time to Recurrence (Years)
E
strata (GEO)
8p11.22–p11.23 (Bicluster 25)
Rest
Percentage of Patients Without Recurrence
Time to Recurrence (Years)
F
ERLIN2
PPAPDC1B
BAG4
TM2D2
DDHD2
FGFR1
LSM1
PROSC
EIF4EBP1
RAB11FIP1
WHSC1L1
ASH2L
BRF2
G
strata (METABRIC)
17q22–q23.3 (Bicluster 33)
Rest
Percentage of Patients Without Recurrence
Time to Recurrence (Years)
H
strata (GEO)
17q22–q23.3 (Bicluster 51)
Rest
Percentage of Patients Without Recurrence
Time to Recurrence (Years)
I
RAD51C
NOL11
SKA2
PPM1D
MRPS23
TUBD1
TRIM37
C17orf71
TLK2
SUPT4H1
CLTC
PTRH2
HEATR6
COIL
RPS6KB1
INTS2
MTMR4
DHX40
RNFT1
APPBP2
VMP1
USP32
DDX5

Figure 6


Hierarchical clustering of biclusters in TCGA
Hierarchical clustering of biclusters in METABRIC
Neurotransmitter secretion and synapse
Mitochondrial translation and respiration + Copy number gains
Copy number gains at 8q24.3
Copy number gains at 17q12 (Her2 Amplicon)
Copy number gains at 8p11.21-p11.23
Cellular division/mitosis + Copy number gains
Response to hormone stimulus, angiogenesis, vasculogenesis
Cell adhesion, extracellular matrix
Immune response
rRNA processing and translation
Mostly copy number gains
No functional association or copy number change